International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 4, Issue 3, Jun 2014, 119-128
© TJPRC Pvt. Ltd.

TRANS STELLAR
• Journal Publications • Research Consultancy

DATA ORIENTED PERCEPTUAL DISCONTINUITY SMOOTHING ALGORITHM

CHANDRA SEKHRAM BONDU[1] & RAMAKRISHNA S[2]

[1]System Analyst, Department of Computer Science, Rashtriya Sanskrit Vidyapeetha, Tirupati, Andhra Pradesh, India

[2]Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India

ABSTRACT

The major drawback of concatenation based Text to Speech Synthesis is the presence of audible discontinuities in the continuous synthetic speech output. The speech units used to generate synthetic speech was extracted from continuous natural utterances. The way of segmenting and extracting the speech units from natural utterances for concatenative speech synthesis has significant effect on the naturalness and intelligibility of synthetic speech. Phoneme, Diphone and Syllable based units are three kinds of popular speech segmentation approaches. Each segmentation approach has its own merits and demerits. Diphone based units are segmented such a way that the transition effect that exists between two phonemes is not lost. When syllable like or phoneme based speech segments were used for generating synthetic speech there were audible disturbances relating to lacking of transitional effect between two phonemes. A Listener can subjectively distinguish an audible difference between the natural utterance and synthetic speech. A method called Perceptual Discontinuity smoothing to reduce such discontinuities using a command line software tool known as sox was experimented. This work majorly focused on subjectively measuring such discontinuities by conducting perceptual listener tests and developing a method to reduce such discontinuities. This work is a part of developing a Text to Speech Synthesis for Sanskrit.

KEYWORDS: Coarticulation Effect, LISS, RISS, TR, MOS, Prosodic Features, Sox, Abhinidhana

INTRODUCTION

Concatenation based Text to Speech Synthesis systems produce synthetic speech by concatenating segments of pre-recorded natural utterances. Out of several approaches tried by the previous researchers, the popular ways of segmenting the natural speech into small units are phoneme, diphone and syllable based. Phonemes are mostly used in concatenative speech synthesis since they are natural linguistic units [1]. Number of phonemes required in the case of Sanskrit is sixty three. Diphone is an utterance that is spread between two phoneme middle points. Diphone units preserve the transition effect that may exist between two phonemes. Syllable units refer to akshara of Sanskrit language. These syllable units are linguistically defined by the phonetics of the particular language. As for as sānskrit phoneticians are concerned, they prefer to adhere to rules enumerated in phonetic Texts [2]. Choosing between the units of speech segments is an area of research. For the languages of India, which exhibit phonemic orthography the characters are best choice for basic units [3]. Irrespective of the unit or nature of the unit of concatenation the synthetic speech output resulted in audible disturbance. The present study is limited to syllable based units of speech used for concatenation. This study was based on assumption that a part of audible disturbances experienced in concatenative synthetic speech was because of lack of co-articulation effect. As mentioned diphone based concatenation preserves the transition between the two phonemes was smooth.

There are several types of discontinuities in synthetic speech generated by concatenation speech synthesis – voice quality, energy, pitch, and formant. It is quite difficult to correlate particular type of discontinuity to connected audible disturbance. All these features of a sound are interdependent and hence cannot be isolated individually for correlating with any specific audible disturbance in synthetic speech. It was mentioned in [4] that the frequency of discontinuities in concatenative synthetic voice depends on many aspects like: 1. variations in speaker's voice, 2. database coverage and 3. change in the recording conditions, 4. nature of the segment unit (phoneme/diphone/syllable) and 5.quality of segmented speech unit. The study of this paper is limited to smoothing discontinuities in synthetic speech arising out of aspects mentioned in point 4 and 5.

RELATED STUDY

The problem of smoothing discontinuities present in synthetic speech is being handled in parametric domain without connecting it to perceptual models. Limited research resources are found which studied to correlate between perceived discontinuity and parametrical spectral features. Perceptual models are efficient in providing overall understanding of discontinuities that may occur in artificial continuous speech, making it simple to learn the sources of an audible discontinuity. It is essential to provide proper correlation between spectral features (formant, energy, and pitch) of sound and perceived discontinuity. Several signal concatenating algorithms are proposed for reducing the discontinuities at spectrum level. To reduce the discontinuities, in [4] an approach called True Period Multi-Band re-synthesis overlap add (TP-MBROLA) is proposed diphone-based synthesis. It is very difficult to avoid spectral discontinuities at concatenation boundaries even though huge quantities of speech segments are available in speech corpus [5].

Several studies tried to establish some correlation between spectral features and audible discontinuity in synthetic speech. Before taking up spectral smoothing it is very essential to determine whether smoothing at a particular junction is required or not. This task can be accomplished by spectral discontinuity detection algorithms. Spectral discontinuity detection algorithms scan the given acoustic signal for abrupt difference in the spectral features. Spectral Distance measuring functions usually determine smoothing requirement at concatenation boundaries. Spectral continuity measures used by previous researchers are: 1. Euclidean distance on log PSP calculated from short time Fast Fourier Transform (FFT) 2. Symmetrical Kullback-Leibler distance on power normalized spectra calculated from short time FFT and LPC [5]. Pick the best modify the least is the recent trend in Corpus based concatenative speech synthesis. It is ideal to desire perfectly matching segment, but practically the same effect has to be achieved with available data. The whole task of spectral smoothing revolves around modifying the three prosodic characteristics –power, pitch and duration.

By modifying these three prosodic characteristics it becomes possible to create wider range of speech segments for concatenation [6]. Spectral smoothing becomes desired when there are situations where contiguous speech segments exhibit perceptually different spectra at their concatenation boundaries. Spectral smoothing and Spectral interpolation are two terms used in relation to removal of spectral discontinuities. Previous researches proved that smooth changes in frequency are perceived as changes with in single speaker, whereas sudden changes in are heard as being change in speaker [7]. In [6] all spectral smoothing techniques are comparatively presented. The basis for this work is derived from studies revealed in [6]. A few researchers attempted to study the manner of articulation with acoustic correlates. Some research studies stressed the connection between manner of articulation and acoustics [9,10]. This work used this idea to reduce the discontinuities that occurred at concatenation boundaries in synthetic speech.

Sox

It stands for Sound eXchange, the Swiss Army knife of audio manipulation is command line audio editing software. This software was used to modify prosodic features of the speech segments. It supports all popular audio formats. This command line sound editing tool can be used to combine any number audio files into one. The features available with sox allow to apply several effects like: balancing the amplitudes of each segment, dynamically compresses with companding effect, change the sampling rates of audio files. SoX's input joiner allows configuring to join multiple audio files using several methods: 1. sequence, 2. concatenate, 3. mix, 4. mix-power, 4.multipy or merge. This command line tool can also be used to trim audio files, segment audio file at specified time points, apply all pass, band pass, band reject, bend, fade-in, fade-out modify pitch, equalize gain at desired frequency, reduce noise. These are few effects that are supported by SoX audio tool. This can be freely downloaded at website: http://sox.sourceforege.net/sox.html [8]. As this command line tool is providing support for all kinds audio effect without much signal processing work, it was used in incorporating/modifying several desired effects on candidate units of synthetic speech to reduce the discontinuities that occurred at concatenation boundaries.

Co-Articulation Effect

It is established that articulation effects on movement of formant is the main cause for the need for spectral smoothing. This effect is more at concatenation when syllable based segments are employed in generating synthetic speech. This effect is more when generative triphonic consonantal syllabic clusters. The co-articuation phenomenon includes changes in articulation and acoustics of phonemes due to the phonetic context. While speaker wishes to utter sequence of phonemes, in a hastiness to utter next phoneme speaker may skip or modify the a part articulation process of current phoneme leads to an effect called co-articulation. As for syllable based units are concerned these articulation effects can be categorized into four kinds of transition: 1.Vowel to Vowel, 2.Vowel to Consonant, 3. Consonant to consonant and 4.consonant to vowel. VC, CV, VV transitions are characterized by formant transitions. The coarticulation effect in CC patter is more complex.

Proposed Algorithm

The proposed approach to reduce the discontinuity at the concatenation boundary has a basis that lack of co-articulation effect at the adjoining boundary has contributes a lot to discontinuity in the synthetic speech. It is essential to incorporate co-articulation base in Text to Speech Synthesis System in order to smoothen discontinuities. While in the process of realizing each utterance in phoneme sequence, the speaker unintentionally tends to modify the articulation process because either his urge to utter next phoneme or lethargy in uttering the current phoneme. This phenomenon is called coarticulation effect. The concept is explained by Sanskrit phonetics in terms of abhinidhana. This algorithm takes two utterances Left-Isolated Speech Segment (LISS) and Right Isolated Speech Segment (RISS) from phoneme sequence, inspects for nature of transition exists between them. It was noticed from extensive study of formant, pitch, intensity contour analysis 5 to 10ms duration that exists between two phonemes is representing the transition effect. These transition segments are also extracted. These transition segments corresponding to respective phonetics transition was padded in between isolated speech segments. While doing so LISS, RISS were also given fade-out and fade-in effects. The duration of isolated phoneme considerably big when compared to the duration phonemes in natural utterances. Accordingly duration effects are also incorporated in synthetic speech using sox command tool.

```
  Algorithm: Data Oriented Perceptual Discontinuity Algorithm
  1. Read the output file obtained from Best Possible Syllable Cluster Algorithm
  2. Read for the joints in the file
  3. Determine  the nature of the joint(CV,VC,CC,VV)
  4. if nature of the discontinuity joint   VV then
              Read the Left isolated Speech segment (LISS)
              # fade out the  end frame of  LISS using sox tool)
              Locate file corresponding to VV transition
              Read the VV transition (VVT)
              Read the Right isolated speech Segment (RISS)
              # fade in the beginning frame of RISS using sox tool)
              Join the  three segments ( LISS+VVT+RISS)
      ELSEIF nature of the discontinuity joint   VC then
              Read the Left isolated Speech segment (LISS)
              # fade out the  end frame of  LISS using sox tool)
              Locate file corresponding to VV transition
              Read the VC transition(VCT)
              Read the Right isolated  speech Segment (RISS)
              # fade in the  beginning  frame of RISS using sox tool)
              Join the  three segments ( LISS+VCT+RISS)
  ELSEIF nature of the discontinuity joint   CC then
              Read the Left isolated Speech segment (LISS)
              # fade out the  end frame of  LISS using sox tool)
              Locate file corresponding to VV transition
              Read the CC transition(CCT)
              Read the Right isolated  speech Segment (RISS)
              # fade in the  beginning  frame of RISS using sox tool)
              Join the  three segments ( LISS+CCT+RISS)
ELSEIF nature of the discontinuity joint   CV then
              Read the Left isolated Speech segment (LISS)
              # fade out the  end frame of  LISS using sox tool)
              Locate file corresponding to VV transition
              Read the CV transition(CVT)
              Read the Right isolated  speech Segment (RISS)
              # fade in the  beginning  frame of RISS using sox tool)
              Join the  three segments ( LISS+VCT+RISS)
  ENDIF
Return The Joined Speech Segment
```

Perceptual Tests

To investigate the nature of transition that exists between two phonemes of the phonetic inventory and correlate them with spectral features, a perceptual listening test was conducted. The phonetic inventory consists of natural recordings representing VV, VC, CC and CV patterns. The test stimuli consisted of CC, VC, CV patterns of synthetic speech. Twenty listeners participated in the experiment. The listeners were asked grade the synthetic speech with maximum score of three if the synthetic sound and natural utterance were equal all aspects. The score three was divided among three vairables Left Phoneme (LP), transition phoneme (TR) and Right Phoneme (RP). The listener was asked to grade one for both LP, RP and TR for their audible quality and zero for poor quality. The Mean Opinion Scores obtained for devanagari phoneme classes are shown on table 1.

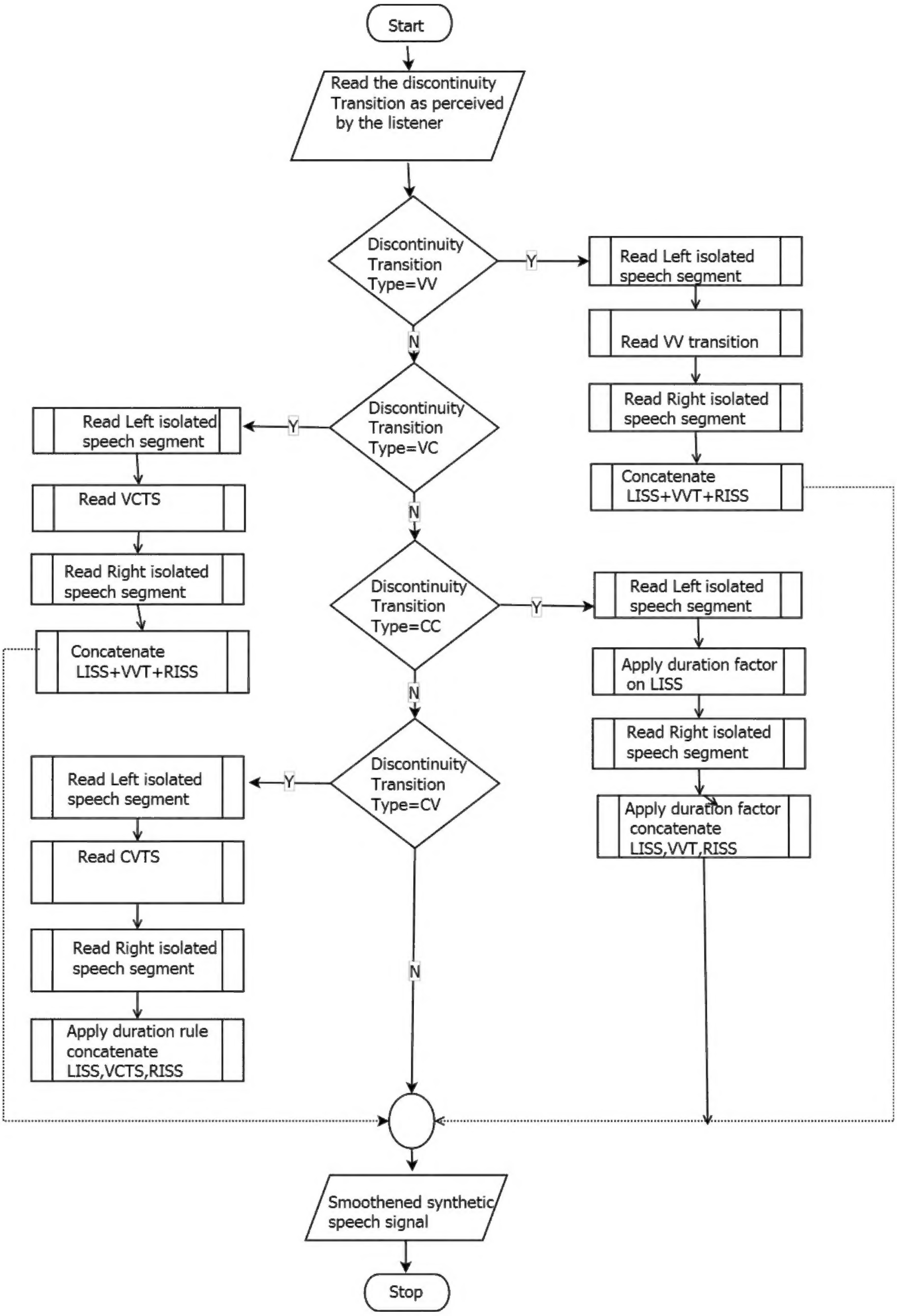

Start
Read the discontinuity Transition as perceived by the listener
Discontinuity Transition Type=VV
Y
N
Read Left isolated speech segment
Read VV transition
Read Right isolated speech segment
Concatenate LISS+VVT+RISS
Discontinuity Transition Type=VC
Read Left isolated speech segment
Read VCTS
Read Right isolated speech segment
Concatenate LISS+VVT+RISS
Discontinuity Transition Type=CC
Read Left isolated speech segment
Apply duration factor on LISS
Read Right isolated speech segment
Apply duration factor concatenate LISS,VVT,RISS
Discontinuity Transition Type=CV
Read Left isolated speech segment
Read CVTS
Read Right isolated speech segment
Apply duration rule concatenate LISS,VCTS,RISS
Smoothened synthetic speech signal
Stop

Figure 1: Flow Chart for Data Oriented Perceptual Discontinuities Smoothing Algorithm

RESULTS

The quality of the isolated speech segments were evaluated using perceptual tests. The proposed algorithm called Data Oriented Perceptual Discontinuity algorithm was experimented for reducing discontinuities at concatenation boundaries. The minimum average mean score of synthetic speech obtained were 0.65 and 0 with and without smoothing algorithm. The maximum average mean score of synthetic speech obtained were 0.95 and 0.2 with and without smoothing algorithm.

CONCLUSIONS

From experiment results data oriented perceptual smoothing algorithm produced impressive results in reducing the discontinuities at concatenation boundaries. This algorithm was tested by integrating with Sanskrit Text to speech Synthesis System. Our perceptual experiments also revealed that segmental quality was also responsible for some of the discontinuities present in synthetic speech.

ACKNOWLEDGEMENTS

The author thanks Mr. Chintalapudi Srinivas, sahitya akademi Prakhya Pavan, Vedam Chandra Sekhar and Mrs. Suvarchala for their valuable support during this work.

REFERENCES

1. Sami Lemmetty, (1999). Review of Speech Synthesis Technology. Masters Thesis, Hensinki University of Technology, Department of Electrical and Communication Engineering
2. Chandra Sekharam Bondu and Rama Krishna S. (2014), "An Approach for Grapheme to Phoneme Alignment for Sanskrit TTS", International Journal of Computer Networking, Wireless and Mobile Communications (IJCNWMC), Vol.4, Issue 2, Apri-2014, 93-100
3. Yegnanarayana, Rajendran S., Ramachandran V R., and Madhukumar A. S. (1994), "Significance of knowledge sources for a text-to-speech system for Indian Languages", Sadhana, Vol. 19, Part 1, February 1994, pp.147-169
4. Shrikant Narayan and Abeer Alwan (2005). Text to Speech Synthesis: New Paradigms and Advances, IMSC Press Multimedia Series, pp.23-38
5. Ju Xu and Lianhong Cai (2008), "Spectral Continuity Measures at Mandarin Syllable Boundaries", Key Laboratory of Pervasive Computing, Ministry of Eduction, Beijing.
6. David T. Chappell, John H. L. Hansen (2002) "A comparison of spectral smoothing methods for segment concatenation based speech synthesis", Speech Communications 36 pp. 343-374
7. Moore BCJ (1997). An Introduction to the Psychology of Hearing, fourth edition. Academic Press, New York.
8. http://sox.sourceforege.net/sox.html
9. Coker, C.H.,(1976) "A model of articulatory dynamics and control", Proc. IEEE 64, pp.452-460
10. Deller Jr., Hansen, J.H.L., Proakis, J.G., (2000), "Discrete Time Processing of Speech Signals", IEEE Press, New York.

APPENDICES

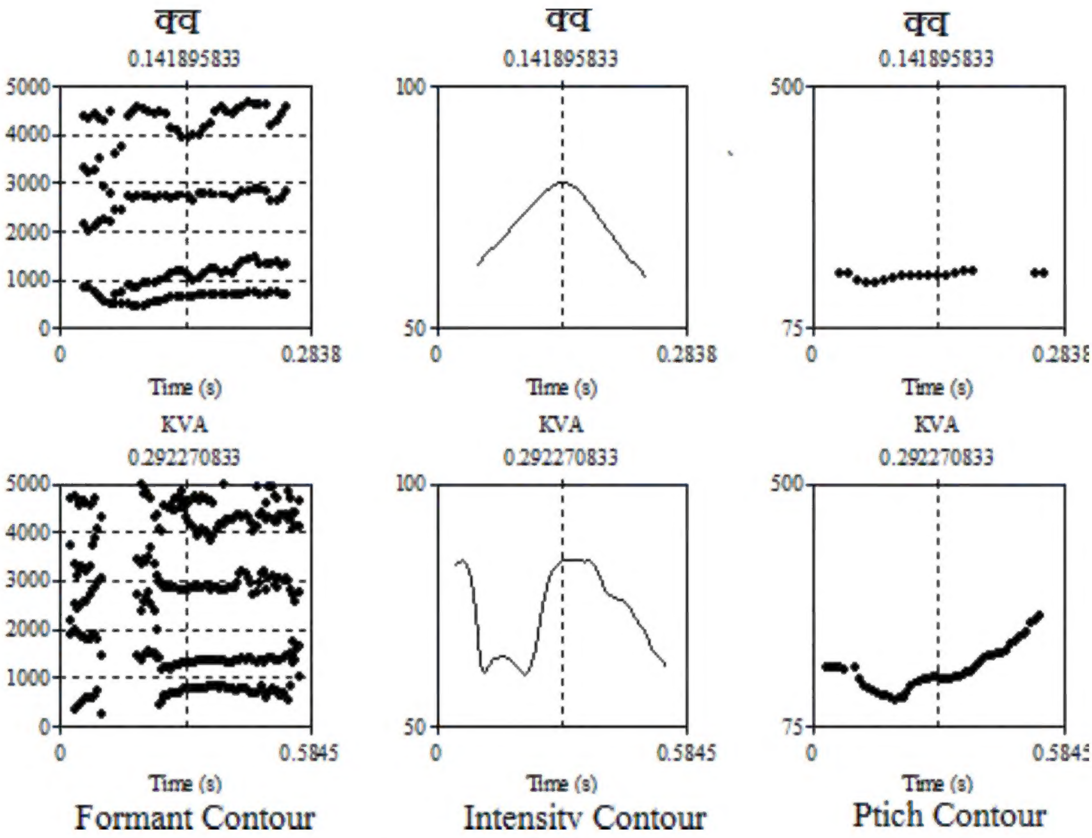
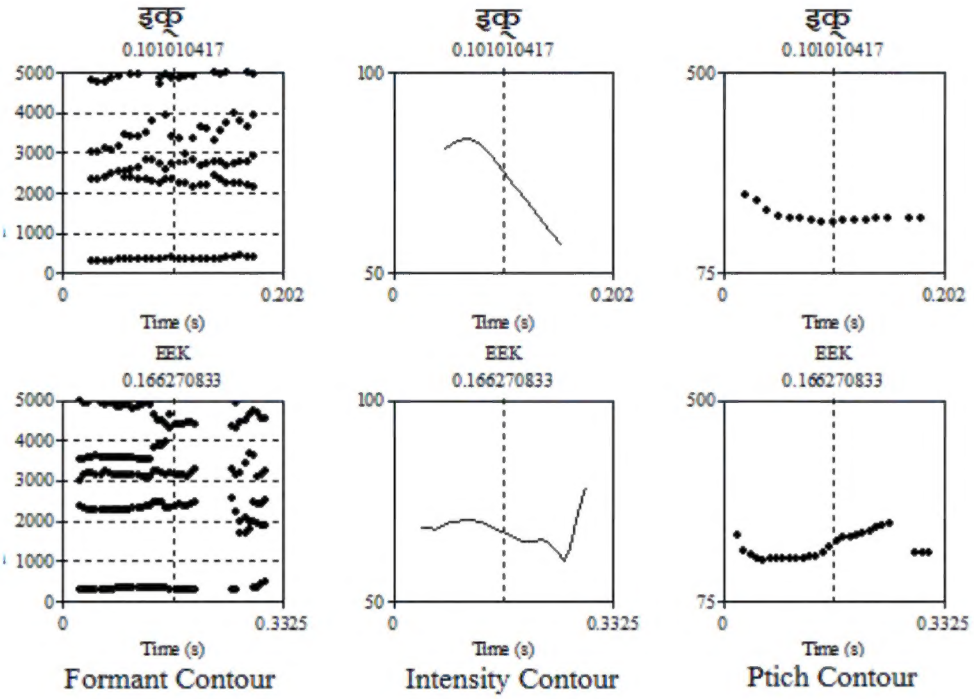


Figure 2: Formant, Intensity and Pitch Contour CV



Pattern

Figure 3: Formant, Intensity and Pitch Contours of VC Pattern

Table 1: Table Showing Mean Opinion Score Obtained from Perceptual Experiment

Phoneme Pair		LP	TR	RP	Total Score	TR after Applying Smoothing
LP	RP					
SPARSA	SPARSA	0.95	0.05	0.95	1.95	0.75
SPARSA	ANUNASIKA	0.95	0.05	0.95	1.95	0.65
SPARSA	SANGHARSHI	0.85	0.05	1.00	1.90	0.85
SPARSA	ANTHASTHA	0.95	0.05	1.00	2.00	0.70
SPARSA	SAMYUKTA	0.95	0.10	0.95	2.00	0.80
SPARSA	SVARA	0.85	0.05	1.00	1.90	0.90
ANUNASIKA	SPARSA	0.75	0.05	0.95	1.75	0.80
ANUNASIKA	ANUNASIKA	0.90	0.10	1.00	2.00	0.95
ANUNASIKA	SANGHARSHI	0.90	0.05	0.95	1.90	0.90
ANUNASIKA	ANTHASTHA	0.85	0.05	0.95	1.85	0.85
ANUNASIKA	SAMYUKTA	0.85	0.0	1.00	1.85	0.75
ANUNASIKA	SVARA	0.95	0.05	0.95	1.95	0.85
SANGHARSHI	SPARSA	0.70	0.20	0.95	1.85	0.85
SANGHARSHI	ANUNASIKA	0.85	0.05	0.95	1.85	0.70
SANGHARSHI	SANGHARSHI	0.90	0.05	0.95	1.90	0.85
SANGHARSHI	ANTHASTHA	0.95	0.15	0.95	2.05	0.95
SANGHARSHI	SAMYUKTA	0.85	0.05	1.00	1.90	0.65
SANGHARSHI	SVARA	0.70	0.05	1.00	1.75	0.80
ANTHASTHA	SPARSA	0.75	0.10	0.85	1.70	0.85
ANTHASTHA	ANUNASIKA	0.85	0.05	0.85	1.75	0.75
ANTHASTHA	SANGHARSHI	0.95	0.05	0.95	1.95	0.90
ANTHASTHA	ANTHASTHA	0.80	0.10	0.95	1.85	0.85
ANTHASTHA	SAMYUKTA	0.90	0.20	1.00	2.10	0.75
ANTHASTHA	SVARA	0.85	0.05	1.00	1.90	0.80
SAMYUKTA	SPARSA	0.85	0.05	0.85	1.75	0.70
SAMYUKTA	ANUNASIKA	0.95	0.10	0.95	2.00	0.75
SAMYUKTA	SANGHARSHI	0.85	0.15	0.85	1.85	0.85
SAMYUKTA	ANTHASTHA	0.75	0.20	0.85	1.80	0.90
SAMYUKTA	SAMYUKTA	0.9	0.10	0.95	1.95	0.65
SAMYUKTA	SVARA	0.9	0.20	0.95	2.05	0.75
SVARA	SPARSA	0.75	0.05	0.85	1.65	0.80
SVARA	ANUNASIKA	0.85	0.05	0.95	1.85	0.80
SVARA	SANGHARSHI	0.95	0.15	0.95	2.05	0.75
SVARA	ANTHASTHA	0.85	0.05	0.90	1.80	0.90
SVARA	SAMYUKTA	0.90	0.15	0.95	2.00	0.90
SVARA	SVARA	1.00	0.05	1.00	2.05	0.85

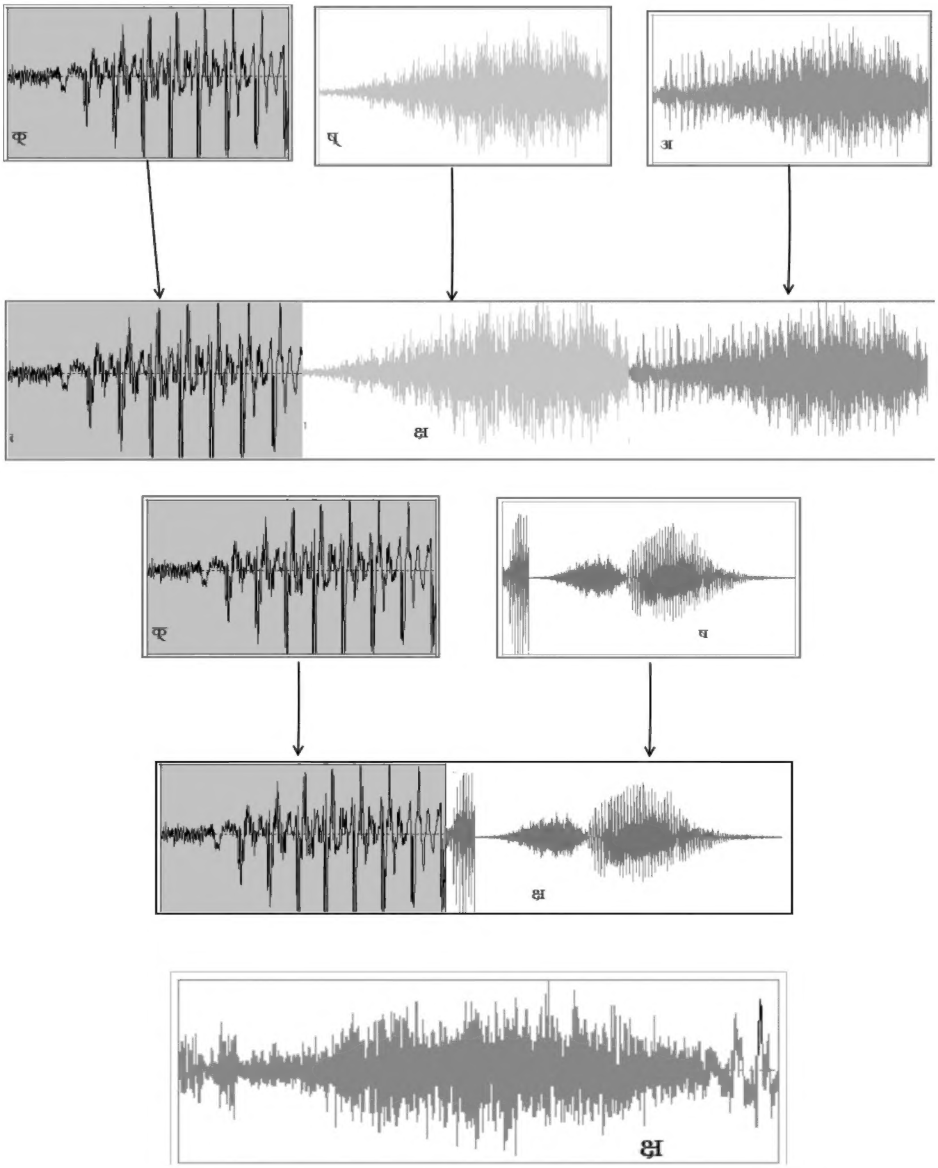


Figure 4: 11Illustrating the Concatenation of Speech Units to Produce Synthetic Speech

Table 2: Table Showing Formant, Intensity and Pitch Contours of Devanagari Phonemes

Letter	Time	Formant Contour				Intesity Contour			Pitch Contour		
Letter	Times	F1_Hz	F2_Hz	F3_Hz	F4_Hz	Intesity Contour dB	Minimum Intensity dB	Maximum Intensity dB	Mean Pitch Hz	Minimum Ptich Hz	Maximum Pitch Hz
अ	0.130708	706.19	1280.14	2812.48	4567.29	71.71	65.12	74.74	160.82	144.42	203.47
इ	0.138375	340.48	2449.46	3214.32	3832.00	68.22	64.86	70.27	190.30	161.79	243.38
उ	0.099187	364.25	867.29	2743.50	4136.43	68.22	64.86	70.27	190.30	161.79	243.38
ऐ	0.136719	381.09	1946.92	2709.80	4258.92	67.70	63.38	69.94	189.61	165.55	213.93
औ	0.154833	471.78	962.10	2922.86	4512.05	72.56	63.07	75.00	150.41	130.57	158.76
क	0.205885	737.25	1322.98	2835.14	3972.34	77.95	54.42	81.95	178.04	159.60	228.82
ग	0.251281	737.18	1382.16	2857.70	4570.20	76.56	60.37	82.95	171.39	125.82	261.16
ङ	0.23974	674.72	1276.28	2710.74	4174.73	75.90	61.74	79.67	147.65	119.86	182.00

Table 2: Cond.,

छ	0.194396	679.94	1384.89	2605.27	3742.39	75.13	60.73	81.78	139.87	80.87	230.11
ज	0.224281	665.16	1657.35	2614.50	3482.95	75.79	63.66	81.42	177.64	123.06	262.93
ट	0.151729	845.30	1369.14	2839.38	3768.18	79.19	64.32	82.29	185.91	164.34	227.35
ड	0.205542	774.93	1317.96	2750.79	3843.27	76.91	64.52	81.91	175.61	121.34	234.91
न	0.176	834.34	1302.63	2761.66	4620.61	75.62	60.61	80.22	75.62	124.50	204.28
त	0.147375	766.65	1384.74	2707.04	4291.95	76.16	69.59	78.67	192.94	173.54	219.67
द	0.227396	816.00	1355.80	2750.04	4357.09	75.73	58.97	80.71	170.20	117.13	231.51
प	0.16199	695.20	1176.05	2498.32	3983.77	74.90	58.86	77.48	188.84	170.48	224.88
ब	0.28224	766.19	1326.05	2940.55	4411.93	76.48	49.94	80.79	162.12	128.20	226.21
म	0.138542	737.58	1053.39	2537.29	4036.91	78.08	67.58	81.13	153.94	128.57	168.48
य	0.31951	700.80	1303.30	2733.87	3942.09	70.17	42.56	75.82	157.99	137.74	212.77
र	0.266594	706.17	1339.96	2488.96	4006.36	73.12	47.95	77.92	169.72	143.67	223.13
ल	0.286677	793.48	1356.69	2868.85	4037.70	78.35	47.94	83.31	170.14	151.17	214.19
व	0.27549	816.98	1348.07	2820.63	3781.75	78.63	53.19	84.50	174.73	122.81	267.06
श	0.287417	737.89	1947.55	2632.51	4638.99	73.30	57.11	77.75	186.23	156.90	246.80
स	0.289177	855.85	1230.29	2716.51	4059.50	76.34	49.69	82.23	204.13	177.15	254.74
ह	0.27624	599.53	1231.29	2732.21	3898.14	68.90	48.91	75.05	178.79	157.33	238.20